Moving beyond the point: An agenda for research in movement analysis with real data

Purves, Ross S ; Laube, Patrick ; Buchin, Maike ; Speckmann, Bettina

DOI: https://doi.org/10.1016/j.compenvurbsys.2014.06.003

Originally published at:
Purves, Ross S; Laube, Patrick; Buchin, Maike; Speckmann, Bettina (2014). Moving beyond the point: An agenda for research in movement analysis with real data. Computers, Environment and Urban Systems, 47:1-4.
DOI: https://doi.org/10.1016/j.compenvurbsys.2014.06.003

Moving beyond the point: an agenda for research in movement analysis with real data

Ross S. Purves[1], Patrick Laube[2], Maike Buchin[3] and Bettina Speckmann[4]

[1] *Department of Geography, University of Zurich, Switzerland*

ross.purves@geo.uzh.ch

[2] *Institute of Natural Resource Sciences, ZHAW Zurich University of Applied Sciences, Switzerland*

patrick.laube@zhaw.ch

[3]*Faculty of Mathematics, Ruhr-Universität Bochum, Germany*

maike.buchin@rub.de

[4]*Department of Mathematics and Computer Science, TU Eindhoven, Netherlands*

b.speckmann@tue.nl

Movement is ubiquitous, whether we are exploring the spread of a disease across the world (Stoddard et al., 2009), the use of bicycles in London (Wood et al., 2011), the behaviour of the West Antarctic Ice Sheet (Rignot et al., 2014), changes in the use of Swiss German dialects (Christen, 1998) or the use of mobile phones in Estonia (Ahas et al., 2007). Despite this ubiquity, GIScience has for many years been accused of a "fetish of the static" (Raper, 2002; Laube et al., 2007), focussing on snapshots and discrete changes, and often treating space and time as independent dimensions (Kucera, 1992). Although many approaches have been proposed to deal with these inadequacies (e.g. Grenon and Smith, 2004; Miller, 2005; Demšar & Virrantaus, 2010), these were often developed in the absence of real, large data sets and based around assumptions rooted in the cartographic, often object focussed world of Geographic Information Systems.

In recent years however a revolution in the potential availability of movement data has brought increasing attention to the limitations of previous approaches to the analysis of movement, and fuelled a step change in the development of methods aiming to deal with spatio-temporal data. Nonetheless, although these methods have been motivated by the claimed availability of very large movement datasets, they have often been developed and tested using small, often pre-filtered and cleaned datasets or simulated data supposed to have similar properties to real movement data, and simplifying assumptions about the properties of data (for example, data with regular and uninterrupted location fixes),

Thus, as authors[1] we have developed and tested algorithms exploring, for instance, football based on 33 seconds of a single football game (Laube et al., 2005), ten caribou (out of a herd consisting of some 100000 individuals) (Laube and Purves, 2006), ten cows moving in a single field over a period of two days (Laube and Purves, 2011), hurricane data for three years featuring a total of 48 storms sampled every six hours (Buchin et al., 2012) or, in the case of algorithmic development in computational geometry sometimes without any reference to real data (e.g. Buchin et al., 2011). This state of affairs is, we would argue, very much representative of the state of the art in research

[1] We deliberately choose to cite our own work here, since our aim is to reveal the disconnect between research (for which we use our own work as exemplars) and, at least potentially, the properties of real movement data.

into movement data, and in particular research based on data collection focussing on fixes directly associated with individual, addressable objects, where the first three Vs of big data (volume, velocity and variety (Laney, 2001; Graham and Shelton, 2013)) are still the exception rather than the rule. Thus, movement research is carried out with trucks and buses in Athens (Vieira et al., 2011), taxis in Shanghai (Lee et al. 2010) or telephone records in the Ivory Coast or Estonia (Ahas et al., 2007; Andrienko et al., 2013), but typically over short periods and with limited integration of multiple datasets.

Therefore, in practice, despite the hype of big data, it appears that the reality, particularly in terms of big data's availability, lags its potential in the realm of movement analysis. There are many reasons for this. For example, in ecology tracking data are very expensive and challenging to collect. This in turn means that animal ecologists typically collect data with a very specific research question in mind, and that tracking data reflect these individual, carefully designed experiments with properties varying according to the research question under investigation. Although data volumes collected have increased significantly, often the variety of these data is sensibly constrained by the particular research question under investigation. Equally, the tracking of humans is certainly a reality, but data are typically held by commercial vendors or governments who must operate under strict privacy legislation restricting their ability to share data (even if this was considered desirable for commercial grounds). Even where such data are released, they are often subject to preprocessing, sometimes significantly modifying data properties (e.g. Andrienko et al., 2013).

None of this should imply that big data won't be of increasing importance in the analysis of movement. However, we contend that there is a need to take a step back, and as big data start to become truly available in many different guises, to consider some of the key issues for the future in the analysis of movement data. Central to these are not only the first three Vs of big data, but also the realities of working with *real data* whose properties are often very different to those we either assume, or encounter in our work with typical pre-prepared datasets. In fact, we would argue that the challenge is not only in dealing with big data *per se*, but rather these real data, warts and all, as collected in a wide variety of applications and with many different aims. We suggest that veracity, validity and volatility, sometimes used to extend the initial 3 Vs of big data to 4, 5 or even 6 Vs, are in fact demonstrably also properties of real data and their quality, irrespective of their volume, the velocity with which they are generated, or the variety of data types and attributes delivered.

In the following we set out a research agenda for movement research, based around the notion of working with real data, typical questions asked of such data, the methods used in analysis and the limitations of the current state of the art. The research agenda draws on the papers in this special issue all of which set out, in different ways, to address issues related to the use of real data in the analysis of movement. Our position is biased by our disciplinary roots (in Geographic Information Science and Computational Geometry), but we would hope that the resulting research agenda provides food for thought for many of those active in this domain. Before setting out this research agenda, we firstly briefly summarise the contributions of the five papers making up the special issue.

The papers

This special issue is made up of five papers, stemming from a workshop held under the auspices of the COST MOVE action (http://move-cost.info/). We briefly summarise the papers here, before embedding them within the research agenda which we set out in the concluding part of this editorial.

Beecham et al. explored check-in data extracted from the London bike hire scheme. Their goal was to analyse commuting behaviour and optimise the use of the scheme. This required solving two classical problems in trajectory data: where are popular places and which routes are used to travel between these? Their approach used methods from visual analytics in a collaborative *chauffeured* analysis approach, where domain experts were key to the workflow demonstrated.

Gudmundsson and Wolle propose a set of automated football analysis tools. As sports movement data increases there is a need for methods capable of analysing such data. The challenge here was to go beyond simple analysis tasks (such as data statistics) to a more meaningful and complex analysis tools (for example analysing player interaction) which are also computationally tractable and of interest to football coaches. The paper demonstrates how approaches from computational geometry can be applied, and constrained by, real world problems in ways which reduce the solution space to more realistic bounds.

Hurter et al. study the problem of extracting contextual information from trajectory data, specifically wind magnitude and direction from recorded aircraft positions derived from radar. They proposed a visual analytics workflow to allow interactive, semi-automatic extraction of these parameters from the data. Importantly, they also compared their approach to data collected independently, and discuss its robustness with respect to data quality and volume.

Meijles et al. study space use by visitors to a natural park. Their goal was to facilitate park management, and the study was carried out in close collaboration with park managers. Having collected data describing visitor behaviour and motivations, error-sensitive analysis was carried out relating behaviour to facilities and signposted trails within the park.

Finally, Perttunen et al. aimed to associate a user's digital identity across multiple datasets. More specifically, their aim was to identify movement traces of the same entity (person or device) captured using independent scanners (in this case using WiFi and Bluetooth). To achieve this, they reformulate the problem as a two-class classification and demonstrate how pruning techniques improve its performance, before discussing privacy implications arising from such approaches.

Sketching a research agenda

In what follows, we set out a research agenda for movement research, focussing on the implications of working with real data, and fuelled by insights derived from the papers making up this special issue. We identify a set of five challenges, ranging from the ways in which we collect data, to those we involve in analysis and the posing of research questions, all of which we believe are of central interest to those involved in research on real movement data, irrespective of the nature of the dataset under investigation.

1) Harness the full diversity of technologies to sense movement and its semantics.

While controlled studies based on a limited number of GPS-tracked individuals will remain an important source for movement data, we should increasingly seize the opportunities offered by the diversity of alternative ways of tracking moving objects. Conventional *Lagrangian* tracking (e.g., GPS receivers producing a sequence of position fixes as an object moves), is increasingly being complemented by *Eulerian* tracking (records of the flow of objects passing by known and fixed check-points or cordons). Be it through mobile phone call detail records (CDR) or commuter journeys emerging from intelligent ticketing systems, the pervasion of technology into our everyday mobility, has resulted in numerous ways of tracking both individuals and summarising aggregate motion. Examples in this special issue include check-in data recording the origins and destinations of bicycles moving within London (Beecham et al.), implying that each individual movement record is made up of exactly two locations, and the use of WiFi and Bluetooth sensors to record passing devices, infer individual movements, and "stitch" these together into individual trajectories, whilst seeking to preserve anonymity (Perttunen et al.). In movement analysis, the need for semantics which go beyond spatio-temporal positions is central. Thus, Meijles et al. collected qualitative data capturing information about a range of parameters including group composition and motivation, using anonymous questionnaires, while Beecham et al. store some basic demographic information associated with individual users of the bicycle hire scheme. The importance of these data, in understanding *why* a particular behaviour took place cannot be understated.

2) *Move beyond the point*

For too long movement research has focussed almost exclusively on the simplistic point-based trajectory conceptual data model for movement. One might argue that it was the simple elegance of the trajectory (in essence a time-stamped sequence of fixes building up a polyline) that contributed to the appeal of movement research in the first place. Studying a moving point is arguably the simplest dynamic spatio-temporal problem and hence a convenient launching point from a static legacy. However, technology has come to rule the conceptual modelling process, with point-based trajectories emerging from GPS systems being the most convenient, and typically undiscussed, option. As is the case in many other areas of GIScience, movement analysis should also seek to allow for more active decisions on the conceptual models used to study both movement and movement spaces, and develop awareness for the implications of such modelling choices. Work with real application problems and domain specialists reveals that many interesting movement problems emerge from adopting an Eulerian perspective and allowing movement to take the form of, for example, flows in fields, check-in sequences, or origin-destination matrices. Hurter et al. take such a perspective in this special issue, using trajectories recording aircraft position in 3D and instantaneous velocities (speed and bearing) to extract wind fields as a continuous field. The potential of working with origin-destination data such as that presented in the special issue by Beecham et al., and the need to use methods which take account the nature of the data in, for example, classifying journey as commutes, also illustrates the importance of moving away from purely point-based representations (in this case classifying work places through a density-based method which was insensitive to outliers but did not return work places as coordinates).

3) A*ccept that the processes and behaviours producing movement cannot be understood from spatio-temporal footprints alone*

Although the above statement seems obvious, a surprising number of studies have sought to understand movement based on trajectories alone. We would submit that true understanding always requires the embedding of the movement in its enabling and limiting context. Figure 1 illustrates this problem through work of the editors, where based on GPS-traces alone we could speculate, but not confirm, why cows moved at velocities far faster than expected, and why they appeared to have a preference for moving within a rectangular region. The integration of movement with context is thus pivotal. This context can come from additional data collected with trajectories as was the case in the work from Beecham et al. and Meijles et al. However, further context can describe the space within which movement can occur (the football pitch in Gudmundsson & Wolle, the street network and key barriers constraining bicycle movement, in Beecham et al. and the nature reserve and path network in Meijles et al). Furthermore, context can also be added by detailed knowledge, and initial hypotheses, about the processes under investigation. Thus, understanding how aircraft behave under prevailing winds allowed Hurter et al. to develop an approach to extracting information on these wind fields. Here, the context is added by domain experts, and we would emphasise how naïve models of process must often be refined with more specialised process and geographical knowledge. A key requirement stemming from this observation is the need for methods which go beyond the most common visual overlay to integrate context, and start to allow for multi-modal, context-aware, integrative movement analysis.

4) *Embrace real but messy movement data*

Stepping away from overly simplifying preliminaries and alibi-validation with small data sets, we see a chance to explicitly advance the toolbox for coping with real but messy movement data. Given GIScience's long history of discussing, documenting and working with uncertainty in purely spatial data, it is truly surprising that method development has often assumed data properties which are rarely present in real data. There is a real need for methods which are not only capable of dealing with coarse, irregular trajectories, but also Eulerian movement data where almost no assumptions about the objects being sensed can be made. Dealing with cordon-based movement data, and in particular its integration with limited information, is the subject Perttunen et al., despite their use of simulated as well as real data, address. Meijles et al. explicitly account for uncertainty and error initially filtering data for unrealistic values, before exploring the robustness of parameters extracted from partial tracks. Hurter et al. discuss the density of sampling points for extracting wind speeds, and use an entropy criterion to evaluate whether sufficient data are available. Finally, Gudmundsson & Wolle explore openness of players to individual football passes, thus extracting possible outcomes at different points in space and time, implicitly expecting that data may reveal not only outcomes, but potential but unrealised results. We see a pressing need for research on movement to more explicitly take account of the nature of real movement data, and associated uncertainties, and thus to develop movement analysis workflows which handle uncertainty on all levels, going beyond simple sensitivity analysis addressing granularity. Equally important is that authors more comprehensively describe data set properties, any preparation that has taken place prior to analysis, and most importantly, that we move towards more reproducible science by making it the norm that movement data associated with papers are placed in the public domain.

5) *Build bridges from abstraction to behaviour (and the question)*

A key experience in movement analysis has been that developing generic analysis tools, as opposed to extracting descriptive parameters (e.g. speed or turning angle), is very challenging. Indeed, we would also argue that even seemingly simple descriptive parameters are often adapted to the nature of the problem and the data under analysis (e.g. can an object take the shortest path between two points, or is it constrained to a network?). Thus, movement analysis which truly aims to understand behaviour must not only take account of the points listed above, but also take the form of a real collaboration between those developing and applying methods, and domain experts. The semantic gap between a flock formalized as a set of points within a circle and the social dynamics within a flock of starlings is simply too wide to be bridged simply through the application of ever more sophisticated movement analysis algorithms and movement mining routines without recourse to experts who actually understand how these birds interact with one another and function as a group. In this special issue we explicitly sought to build this bridge, and we would argue that the resulting papers have gained in relevance through their interaction with specialists. Thus, Gudmundsson & Wolle consulted football coaches, Hurter et al. worked with air traffic controllers and meteorologists, Beecham et al. identified and iteratively answered research questions through interactions with transport specialists and Meijles et al. integrated the state forestry commission into their research team. We firmly believe that the importance of integrating domain knowledge into research on movement, not only to discuss results but also throughout the research process is of central importance. Indeed, as illustrated by some of the papers presented in this special issue, large datasets often require inductive approaches, where the questions which are posed are centrally dependent on domain expertise. Thus, we would argue, a key way in which bridges can be built from computationally sensible, but potentially meaningless abstractions, towards truly context aware analysis of movement, is the consideration of the issues set out in this research agenda, and a corresponding acceptance that the data are unlikely to speak for themselves.

Acknowledgements

The editors of the special issue gratefully acknowledge the support of the COST action MOVE (IC0903) in the research reported here, and the organisation of the workshop and networking activities which made this special issue possible. We also thank Tim Wark, CSIRO ICT Centre and CSIRO Livestock Industries, Rockhampton, Australia, for giving us access to the GPS data illustrated in Figure 1.

References

Ahas, R., Aasa, A., Mark, Ü., Pae, T., & Kull, A. (2007). Seasonal tourism spaces in Estonia: Case study with mobile positioning data. *Tourism management*, *28*(3), 898-910.

Andrienko, G., Andrienko, N., Fuchs, G. (2013). Multi-perspective analysis of D4D fine resolution data. Data for Development (D4D 2013). Available at http://geoanalytics.net/and/d4d2013/

Buchin, K., Buchin, M., Van Kreveld, M., & Luo, J. (2011). Finding long and similar parts of trajectories. *Computational Geometry*, *44*(9), 465-476.

Buchin, M., Dodge, S., & Speckmann, B. (2012). Context-aware similarity of trajectories. In Ningchuan Xiao, Mei-Po Kwan, Michael F. Goodchild, Shashi Shekhar (Eds.): *Geographic Information Science - 7th International Conference*, GIScience 2012, Columbus, OH, USA (43-56). Springer Berlin Heidelberg.

Christen, H. (1998). Convergence and divergence in the Swiss German dialects.*Folia Linguistica*, *32*(1-2), 53-68.

Demšar, U., & Virrantaus, K. (2010). Space–time density of trajectories: exploring spatio-temporal patterns in movement data. *International Journal of Geographical Information Science*, *24*(10), 1527-1542.

Graham, M., & Shelton, T. (2013). Geography and the future of big data, big data and the future of geography. *Dialogues in Human Geography*, *3*(3), 255-261.

Grenon, P., & Smith, B. (2004). SNAP and SPAN: Towards dynamic spatial ontology. *Spatial cognition and computation*, *4*(1), 69-104.

Kucera, G. (1992). *Time in Geographic Information Systems*. CRC Press.

Laney, D. (2001). 3D data management: controlling data volume, velocity, and variety. Application Delivery Strategies, META Group, 6 February, 2001, http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3DData-Management-Controlling-Data-Volume-Velocityand-Variety.pdf

Laube, P., & Purves, R. S. (2006). An approach to evaluating motion pattern detection techniques in spatio-temporal data. *Computers, Environment and Urban Systems*, *30*(3), 347-374.

Laube, P., & Purves, R. S. (2011). How fast is a cow? Cross-Scale Analysis of Movement Data. *Transactions in GIS*, *15*(3), 401-418.

Laube, P., Dennis, T., Forer, P., & Walker, M. (2007). Movement beyond the snapshot–Dynamic analysis of geospatial lifelines. *Computers, Environment and Urban Systems*, *31*(5), 481-501.

Laube, P., Imfeld, S., & Weibel, R. (2005). Discovering relative motion patterns in groups of moving point objects. *International Journal of Geographical Information Science*, *19*(6), 639-668.

Lee, K., Yi, Y., Jeong, J., Won, H., Rhee, I., & Chong, S. (2010, March). Max-contribution: On optimal resource allocation in delay tolerant networks. In *INFOCOM, 2010 Proceedings IEEE* (1-9). IEEE.

Miller, H. J. (2005). What about people in geographic information science. In *Re-Presenting Geographic Information Systems, John Wiley*, 215-242.

Raper, J. (2002). The dimensions of GIS. Keynote at the 2nd International Conference of Geographic Information Science, Boulder, CO, September 25-28.

Rignot, E., Mouginot, J., Morlighem, M., Seroussi, H., & Scheuchl, B. (2014). Widespread, rapid grounding line retreat of Pine Island, Thwaites, Smith, and Kohler glaciers, West Antarctica, from 1992 to 2011. *Geophysical Research Letters*.

Stoddard, S. T., Morrison, A. C., Vazquez-Prokopec, G. M., Soldan, V. P., Kochel, T. J., Kitron, U., Elder, J.P. & Scott, T. W. (2009). The role of human movement in the transmission of vector-borne pathogens. *PLoS Neglected Tropical Diseases*, *3*(7), e481.

Vieira, M. R., Bakalov, P., & Tsotras, V. J. (2011). FlexTrack: A system for querying flexible patterns in trajectory databases. In *Advances in Spatial and Temporal Databases* (475-480). Springer Berlin Heidelberg.

Wood, J., Slingsby, A., & Dykes, J. (2011). Visualizing the dynamics of London's bicycle-hire scheme. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *46*(4), 239-251.

List of figures

Figure1: A GPS trajectory with no additional information (top left) reveals little about the behaviour of the object of interest, other than that they reach unlikely velocities of 75 km/h. Adding some basic context (bottom left) in the form of a background map shows that these high velocities occur on a road, suggesting the cows are being transported. Zooming in to detailed GPS fixes (right) with high resolution imagery shows that the cows have a preference for the, physically observable, edges of the field, which also appears to be subdivided in the middle. Background mapping ©OpenStreetMap and contributors and ESRI and contributors.